Method for Computing Protein Binding Affinity

Charles F. F. Karney[*] and Jason E. Ferrara
Sarnoff Corporation, Princeton, NJ 08543-5300

Stephan Brunner[†]
Locus Pharmaceuticals, Inc., Blue Bell, PA 19422-2700
(Dated: July 14, 2004; revised September 9, 2004)

A Monte Carlo method is given to compute the binding affinity of a ligand to a protein. The method involves extending configuration space by a discrete variable indicating whether the ligand is bound to the protein and a special Monte Carlo move which allows transitions between the unbound and bound states. Provided that an accurate protein structure is given, that the protein-ligand binding site is known, and that an accurate chemical force field together with a continuum solvation model is used, this method provides a quantitative estimate of the free energy of binding.

Keywords: free energy; binding affinity; Monte Carlo methods; equilibrium constants; proteins

Introduction

Many drugs work by binding to a target protein in an organism and affecting the action of this protein. The binding of the drug molecule, the ligand L, to the protein P under physiological conditions is usually reversible (characterized by weak chemical interactions rather than covalent bonds),
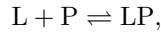
$$L + P \rightleftharpoons LP,$$

and, in equilibrium and in the dilute limit, the concentration of the ligand-protein complex [LP] is given by the dissociation constant

$$K_d = \frac{[L][P]}{[LP]}.$$

It is convenient to define the binding affinity as

$$pK_d = -\log_{10}\left(\frac{K_d/N_A}{1 \,\mathrm{kmol\,m^{-3}}}\right),$$

where N_A is the Avogadro constant. A high value for pK_d is crucial to obtaining a good drug molecule and, consequently, the ability to compute pK_d accurately would greatly accelerate drug discovery by allowing many molecules to be screened *in silico* before any time-consuming syntheses and assays are done. The dissociation constant can also be expressed in terms of the binding free energy ΔF,

$$K_d = \exp(\beta \Delta F)/V_0, \tag{1}$$

where V_0 is the system volume, $\beta = 1/(kT)$, T is the temperature, and k is the Boltzmann constant. Similarly pK_d is given by

$$pK_d = -\frac{\beta \Delta F}{\ln 10} + \log_{10}(V_0 N_A \times 1 \,\mathrm{kmol\,m^{-3}}).$$

The quantity ΔF is the free energy difference of the ligand and the protein forming a bound complex LP, the "bound system," compared to the ligand and the protein isolated from one another L + P, the "unbound system."

In order to compute pK_d, we require (1) a sufficiently accurate model of the protein and the ligand and their interaction and (2) a good way to compute the resulting value of pK_d. In this paper, we assume the first requirement is fulfilled and instead focus on meeting the second.

The standard methods of computing free energies [1, 2] are not capable of computing ΔF directly because the unbound and bound systems are too dissimilar, which hinders transitions between these systems. Instead, typically, two close ligands L_a and L_b, are compared separately unbound and bound to the protein, thereby obtaining the difference in the free energies $\Delta F_a - \Delta F_b$.

We present here a practical method for directly computing ΔF and hence pK_d. The method consists of: (1) formulating the problem in an extended phase space which allows the unbound and bound systems to be treated as a single system and K_d to be expressed as the ratio of two canonical averages; (2) introducing a new Monte Carlo move, the "wormhole move," to make transitions between the unbound and bound states in this extended system; and (3) a method to determine the "portals" needed for the wormhole move.

Formulation

Consider a system of volume V_0 consisting of a ligand molecule L and a protein molecule P dissolved in N_S molecules S. The overall state of the system is given by $[\Gamma, \Gamma_S]$ where Γ represents the phase space configuration of L and P and Γ_S represents the configurations of all the solvent molecules S. The energy of the system is given by $E(\Gamma, \Gamma_S)$ and, in equilibrium, the system obeys the Boltzmann distribution [3]

$$f(\Gamma, \Gamma_S) = \frac{\exp[-\beta E(\Gamma, \Gamma_S)]}{\int \exp[-\beta E(\Gamma, \Gamma_S)]\, d\Gamma\, d\Gamma_S}.$$

It is frequently useful to average over the configurations of the solvent molecules by integrating the Boltzmann distribution

over Γ_S to give a reduced Boltzmann distribution

$$
\begin{aligned}
f(\Gamma) &= \int f(\Gamma, \Gamma_\mathrm{S})\, d\Gamma_\mathrm{S} \\
&= \frac{\exp[-\beta E(\Gamma)]}{\int \exp[-\beta E(\Gamma)]\, d\Gamma},
\end{aligned}
$$

where

$$
E(\Gamma) = -\frac{1}{\beta} \ln\left(\int \exp[-\beta E(\Gamma, \Gamma_\mathrm{S})]\, d\Gamma_\mathrm{S} \right)
$$

is the energy of the system with ligand and protein configurations specified by Γ and with the equilibrium effects of the solvent implicitly included as a solvation free energy. In this paper, we will assume that $E(\Gamma)$ is given.

Typical molecular interactions have a short range. In view of this, let us define Σ_0 to represent all accessible Γ space (i.e., L and P somewhere in the system volume V_0), and Σ_1 to represent that portion of Σ_0 where there is an appreciable interaction between L and P which therefore form the complex LP. In the phase-space volume Σ_1 we write the energy as $E_1(\Gamma)$ which is just an alternate notation for the full energy $E(\Gamma)$, while in the volume $\Sigma_0 - \Sigma_1$ we may write the energy as $E_0(\Gamma)$ which we define as the "unbound" energy, i.e., $E(\Gamma)$ excluding the interaction between L and P. The dissociation constant can then be written as

$$
\begin{aligned}
K_d &= \frac{1}{V_0} \frac{\left(\int_{\Sigma_0 - \Sigma_1} e^{-\beta E(\Gamma)}\, d\Gamma\right)^2}{\int_{\Sigma_0} e^{-\beta E(\Gamma)}\, d\Gamma \int_{\Sigma_1} e^{-\beta E(\Gamma)}\, d\Gamma} \\
&= \frac{1}{V_0} \frac{\left(\int_{\Sigma_0 - \Sigma_1} e^{-\beta E_0(\Gamma)}\, d\Gamma\right)^2}{\int_{\Sigma_1} e^{-\beta E_1(\Gamma)}\, d\Gamma} \times \\
&\quad \frac{1}{\int_{\Sigma_0 - \Sigma_1} e^{-\beta E_0(\Gamma)}\, d\Gamma + \int_{\Sigma_1} e^{-\beta E_1(\Gamma)}\, d\Gamma}.
\end{aligned}
$$

In the dilute limit $V_0 \to \infty$, this can be simplified by ignoring the second term in the denominator of the last factor and by extending the limits of the integrals over $\Sigma_0 - \Sigma_1$ to include Σ_1. In extending the definition of $E_0(\Gamma)$ to $\Gamma \in \Sigma_1$, we include the intramolecular energy and the solvation free energy but continue to omit the intermolecular (ligand-protein) energy. This yields [1, 4, 5]

$$
K_d = \frac{1}{V_0} \frac{\int_{\Sigma_0} \exp[-\beta E_0(\Gamma)]\, d\Gamma}{\int_{\Sigma_1} \exp[-\beta E_1(\Gamma)]\, d\Gamma}. \tag{2}
$$

The Helmholtz free energy of the unbound and bound systems is [3]

$$
F_\lambda = -\frac{1}{\beta} \ln\left(\int_{\Sigma_\lambda} \exp[-\beta E_\lambda(\Gamma)]\, d\Gamma \right),
$$

for $\lambda = 0$ and 1, and Eq. (1) is obtained from Eq. (2) with $\Delta F = F_1 - F_0$. The definition of K_d, Eq. (2), is strictly independent of V_0 because of translational symmetry (ignoring

boundary effects). It is also independent of the precise definition of Σ_1, provided that Σ_1 includes the protein-ligand binding site and does not include a "macroscopic" volume beyond this.

In this formulation, we have assumed that the system volume V_0 is fixed. However, in most physiological systems, the pressure is held constant and the binding affinity is then related to the differences in the Gibbs free energy which introduces a correction term which is the product of the pressure and the change in the volume caused by the formation of the LP complex [6]. We expect this correction to be small for typical ligand-protein interactions in solution.

We would like to cast Eq. (2) as the ratio of canonical averages which can be computed using the canonical-ensemble Monte Carlo method [7]. To achieve this, we combine the unbound and bound systems by extending phase space to include a discrete index $\lambda \in \{0, 1\}$ and consider a system in this extended space with energy $E_\lambda^*(\Gamma)$ for which the canonical average is defined by

$$
\langle X \rangle = \frac{\sum_\lambda \int d\Gamma\, \exp[-\beta E_\lambda^*(\Gamma)] X_\lambda(\Gamma)}{\sum_\lambda \int d\Gamma\, \exp[-\beta E_\lambda^*(\Gamma)]}.
$$

We take $E_\lambda^*(\Gamma)$ to be infinite for $\Gamma \notin \Sigma_\lambda$ and finite otherwise. Now Eq. (2) can be rewritten as

$$
K_d = \frac{1}{V_0} \frac{\left\langle \delta_{\lambda 0} \exp\left(-\beta[E_0(\Gamma) - E_0^*(\Gamma)]\right) \right\rangle}{\left\langle \delta_{\lambda 1} \exp\left(-\beta[E_1(\Gamma) - E_1^*(\Gamma)]\right) \right\rangle}, \tag{3}
$$

where $\delta_{\lambda\mu}$ is the Kronecker delta. If $E_\lambda^*(\Gamma) \approx E_\lambda(\Gamma)$, the terms being averaged are $O(1)$. Because the definition of K_d is independent of V_0, we can pick $V_0 \sim 1/K_d$ so that approximately the same number of samples contribute to each of the canonical averages. We show later, Eq. 7, that this choice minimizes the error in the estimate of K_d.

We can evaluate $E_\lambda^*(\Gamma)$ with short energy cutoffs allowing it to be computed more quickly than $E_\lambda(\Gamma)$ and the terms contributing to the averages in Eq. (3) can be accumulated every hundred steps, for example. Since there is typically a high correlation between successive steps in a Monte Carlo simulation, this method allows the averages to be computed to a given degree of accuracy in less time than if we had used $E_\lambda^*(\Gamma) = E_\lambda(\Gamma)$.

The extension of phase space has been used in other free energy calculations, to combine, for example, systems at several different temperatures [8] or to treat the "reaction coordinate" controlling the transition between two chemical species as a dynamic variable [9, 10]. In our case, the use of the wormhole Monte Carlo (described in the next section) allows us to include just the two systems of interest without the need to compute the properties of (possibly unphysical) intermediate systems.

Wormhole Monte Carlo

We can compute the canonical averages in Eq. (3) using the Monte Carlo method [7] to make steps from $[\Gamma, \lambda]$ to $[\Gamma', \lambda']$
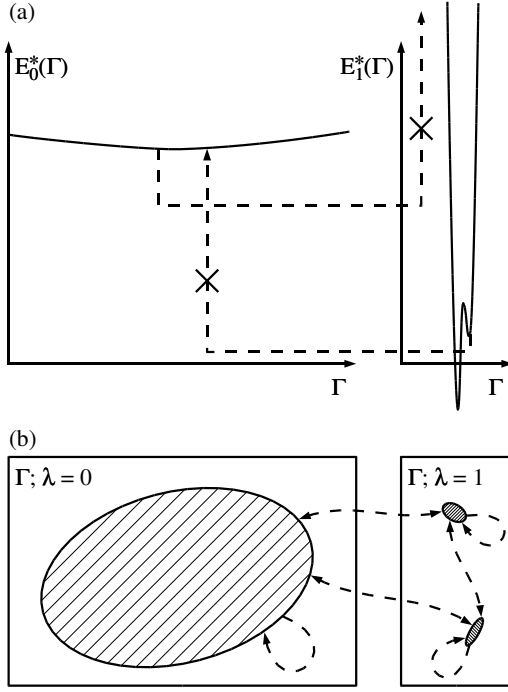
FIG. 1: (a) Schematic representation of $E_0^*(\Gamma)$ (shallow and wide) and $E_1^*(\Gamma)$ (deep and narrow). Conventional Monte Carlo moves between $\lambda = 0$ and 1 (shown as dashed lines) are nearly always rejected because they lead to large increases in energy. (b) Schematic representation of typical portals for the wormhole moves for the case illustrated in (a), The large ratio of the volume of the unbound ($\lambda = 0$) portal compared to the bound ($\lambda = 1$) portals compensates for the higher energy of the unbound configurations. This results in accepted wormhole moves (dashed lines) between all the portals.

with probability

$$\min\big[1, \exp\big(-\beta[E_{\lambda'}^*(\Gamma') - E_\lambda^*(\Gamma)]\big)\big].$$

However, the estimate of the ratio in Eq. (3) will be very poor, because transitions between $\lambda = 0$ and 1 will be extremely rare—typically, E_0^* is shallow and wide, while E_1^* is deep and narrow; see Fig. 1(a). One possible way of remedying this is to treat λ as a continuous variable [9, 10], providing a suitably interpolated definition of $E_\lambda(\Gamma)$, and allowing Monte Carlo steps with small changes in λ. In practice, this approach only "works" if the two endpoints are sufficiently similar, thus limiting this method to the comparison of chemically close molecules.

Here we propose an alternative way of carrying out a Monte Carlo simulation of the $E_\lambda^*(\Gamma)$ system. We restrict the standard moves to changes in Γ only, and allow changes in λ via "wormhole moves" [11] which connect otherwise disconnected regions of configuration space. This obviates the need to treat (possibly unphysical) intermediate values of λ, permitting us to compute the free energy differences needed to determine the absolute binding affinity.

Assume we have some system defined on a phase space Υ

whose equilibrium distribution is proportional to $g(\Upsilon)$. The canonical average of a quantity $X(\Upsilon)$ is defined by

$$\langle X \rangle = \frac{\int d\Upsilon\, g(\Upsilon) X(\Upsilon)}{\int d\Upsilon\, g(\Upsilon)}.$$

In our application, we make the identification $\Upsilon = [\Gamma, \lambda]$, $\int d\Upsilon = \sum_\lambda \int d\Gamma$, and $g(\Upsilon) = \exp[-\beta E_\lambda(\Gamma)]$.

Let us define a set of "portal functions," $w, w', w'', \ldots$, on Υ, with properties

$$0 \le w(\Upsilon) \le 1/v < \infty,$$

$$\int d\Upsilon\, w(\Upsilon) = 1,$$

where v is a representative phase-space volume of the portal function. A wormhole move consists of the following steps: select a pair of portals (w, w') with probability $p_{ww'}$; reject the move with probability $1 - vw(\Upsilon)$, where Υ is the current state; otherwise, with probability $vw(\Upsilon)$, pick a configuration Υ' with probability $w'(\Upsilon')$; and accept the move to Υ' with probability $P_{ww'}(\Upsilon, \Upsilon')$. If the move is rejected, Υ is retained as the new state. We determine $P_{ww'}(\Upsilon, \Upsilon')$ by demanding that the rate of making transitions from Υ to Υ' via portals (w, w') is balanced by the reverse rate from Υ' to Υ via portals (w', w), i.e.,

$$R_{ww'}(\Upsilon, \Upsilon') = R_{w'w}(\Upsilon', \Upsilon)$$

(the condition of detailed balance), where the rates are given by

$$R_{ww'}(\Upsilon, \Upsilon') = p_{ww'} \frac{g(\Upsilon)}{\int d\Upsilon\, g(\Upsilon)} \times \\ vw(\Upsilon)w'(\Upsilon')P_{ww'}(\Upsilon, \Upsilon').$$

A possible solution for the acceptance probability is

$$P_{ww'}(\Upsilon, \Upsilon') = \min\left(1, \frac{p_{w'w}}{p_{ww'}} \frac{g(\Upsilon')}{g(\Upsilon)} \frac{v'}{v}\right),$$

where we have made use of the identity $\alpha \min(1, \alpha^{-1}) = \min(1, \alpha)$, which is valid for $\alpha > 0$.

In order to apply this move, let us use a specific rather simple form for the portal functions, $w(\Upsilon)$, namely

$$w(\Upsilon) = \begin{cases} 1/v, & \text{for } \Upsilon \in w, \\ 0, & \text{otherwise,} \end{cases}$$

where we now denote a portal by w which defines an arbitrary subset of Υ space of volume v. In principle, there is no restriction on the choice of the portals; however, practical considerations, discussed below, dictate how they are chosen. Furthermore, for simplicity, we will assume that the wormhole probabilities are all equal, $p_{ww'} = \text{const}$.

Let us now describe the wormhole move in $[\Gamma, \lambda]$ space. Starting with a configuration $[\Gamma, \lambda]$, first pick a random portal

w. If $[\Gamma, \lambda] \notin w$, then reject the move. Otherwise, pick a random configuration $[\Gamma', \lambda']$ uniformly in a randomly chosen portal w', and accept the move to $[\Gamma', \lambda']$ with probability

$$\min\left(1, \frac{\exp[-\beta E_{\lambda'}^*(\Gamma')]}{\exp[-\beta E_{\lambda}^*(\Gamma)]} \frac{v'}{v}\right). \tag{4}$$

This differs from the standard Boltzmann acceptance probabilty by the ratio of volumes, v'/v, which can be large enough to compensate for the difference in the mean energies in the portals. Indeed, if the mean energy of configurations in w scales as $\beta^{-1} \ln v + \text{const.}$, the acceptance probability, Eq. (4), is $O(1)$, thereby allowing moves between shallow wide wells and deep narrow ones; see Fig. 1(b). In practice, each portal will occupy one of the energy wells of either $E_0^*(\Gamma)$ or $E_1^*(\Gamma)$. This implies that the $\lambda = 0$ portals will permit unrestricted translation and rotation of the ligand and the protein subject to the V_0 constraint, and the $\lambda = 1$ portals will allow unrestricted motion of one of the molecules.

The wormhole move embodies two concepts which have been used separately in other works: stretching or shrinking Γ space when making the move compensates for possibly large energy differences between wells [12, 13]; and jumping between disconnected regions of phase space enables the Markov chain to explore regions of phase space separated by large energy barriers [14, 15]. In the following sections, we will show how these elements may be combined to permit the computation of protein binding affinities with realistic force fields.

Finding the portals

In order for the wormhole method to be practical, we need a reliable way of choosing the portals. We describe this process first in the general case. Let us write $\Gamma = [\Gamma_L, \Gamma_P]$, where $\Gamma_M = [X_M, \Xi_M]$ represents the state of molecule M, X_M represents its position and orientation, and Ξ_M represents its conformation.

We assume that Ξ_M is expressed in such a way that any constraints on the positions of the atoms in M (e.g., bond lengths and bond angles) is implicitly accounted for, so that the dimensionality of Ξ_M reflects the number of degrees of conformational freedom, n_M, for this molecule.

Because of the translational and rotational symmetry of the system, certain components of Γ are ignorable. We can therefore write $E_{\lambda}^*(\Gamma) = E_{\lambda}^*(\Xi_{\lambda})$, with

$$\Xi_0 = [\Xi_L, \Xi_P],$$
$$\Xi_1 = [Y, \Xi_L, \Xi_P],$$

where $Y = X_L - X_P$ denotes the position and orientation of the ligand relative to the protein. The dimensionality of Ξ_{λ} is n_{λ} with $n_0 = n_L + n_P$ and $n_1 = n_0 + 6$.

The strategy for determining the portals is to carry out conventional canonical Monte Carlo simulations with $E_{\lambda}^*(\Xi_{\lambda})$ separately for $\lambda = 0$ and 1. For each λ, we obtain a canonical set of configurations $\{\Xi\}$ (suppressing the λ subscripts for brevity), to which we fit n-dimensional ellipsoids, as follows. First we try to fit a single ellipsoid to $\{\Xi\}$. The center of the ellipsoid is given by the mean configuration $\langle\Xi\rangle$. For each configuration in $\{\Xi\}$, we determine the deviation from the mean, $\delta\Xi = \Xi - \langle\Xi\rangle$, and compute a covariance matrix for the configurations which can be diagonalized as

$$\langle\delta\Xi\,\delta\Xi\rangle = \mathsf{P}\Lambda\mathsf{P}^{\mathrm{T}},$$

where P is the matrix of (column) eigenvectors and Λ is the diagonal matrix of eigenvalues. Because of the properties of the covariance, P is real and orthogonal, $\mathsf{P}^{-1} = \mathsf{P}^{\mathrm{T}}$, and the eigenvalues are real and non-negative. If there are no hidden constraints on the motion, we additionally can assume that the eigenvalues are strictly positive. We find it convenient to define

$$\mathsf{B} = \mathsf{P}\Lambda^{1/2}, \qquad \mathsf{B}^{-1} = \Lambda^{-1/2}\mathsf{P}^{\mathrm{T}},$$

so that we can write

$$\langle\delta\Xi\,\delta\Xi\rangle = \mathsf{B}\mathsf{B}^{\mathrm{T}}.$$

We take the semi-axes of the ellipsoid to be the columns of $\sqrt{n}\mathsf{B}$. The multiplier here, $\sqrt{n}$, is chosen to ensure that $O(1)$ of the configurations in $\{\Xi\}$ lie within the ellipsoid. This choice is motivated by considering a symmetric n-dimensional Gaussian

$$f(\mathbf{r}) = \frac{\exp(-\frac{1}{2}r^2)}{(2\pi)^{n/2}},$$

for which we have

$$\langle r^2\rangle = \int r^2 f(\mathbf{r})\,d^n\mathbf{r} = n.$$

Ellipsoids are a natural choice to use to fit the set of configurations for the following reasons: (1) The iso-density contours of the distribution in a harmonic well are ellipsoids. (2) It is easy to sample points randomly from an ellipsoid. (3) Conversely, it is easy to test that a point lies inside an ellipsoid. (4) The volume of an n-dimensional ellipsoid is given by

$$v_n = \frac{\pi^{n/2}}{(n/2)!}\prod_{i=1}^{n} a_i,$$

where a_i is the length of ith semi-axis. Note well the degenerate case of this result, $v_0 = 1$, which is used if the protein and ligand are both rigid ($n_0 = 0$). The sampling and testing of points, (2) and (3), can either be accomplished by transforming with the matrix B. Alternatively, we can use the simpler Cholesky decomposition of the covariance matrix

$$\langle\delta\Xi\,\delta\Xi\rangle = \mathsf{C}\mathsf{C}^{\mathrm{T}},$$

where C is a lower triangular matrix. Both C and C^{-1} may be computed by direct (non-iterative) methods [16].

Having defined an ellipsoid in this way, we test its suitability as a portal by demanding that $O(1)$ of the configurations sampled uniformly from it have energies close to its mean energy $\langle E^*(\Xi) \rangle$. If this test fails, we split $\{\Xi\}$ into two sets according to the sign of $\delta\Xi$ projected along the largest semi-axis of the ellipsoid and construct new trial portals from each of these sets.

The ellipsoids that result from this process constitute our portals. When computing the volumes of the portals for use in Eq. (4), we need to account for the freedom to place $2 - \lambda$ molecules at arbitrary positions and orientations in the volume V_0. In practice, this means we multiply the volume of the $\lambda = 0$ portals by σV_0 where V_0 is the translational volume and σ is the orientational volume (given below).

In order to complete the specification of the portals, we need to describe how $\langle \Xi \rangle$ and $\delta\Xi$ are formed, since, as a consequence of working in the subspace where the molecular constraints are implicitly satisfied, Ξ is not simply a point in $\mathbb{R}^n$. We wish to represent each ellipsoid on a locally Cartesian space $\mathbb{R}^n$ which lets us use familiar formulas for defining the ellipsoid. We also demand that the mapping to $\mathbb{R}^n$ have constant Jacobian in order to maintain detailed balance when sampling from the ellipsoids.

We illustrate this by considering the case of the protein and the ligand being made up of $n_{\mathrm{M}} + 1$ rigid fragments simply connected by n_{M} bonds each of which allow rotation only (i.e., the bond lengths and bond angles are fixed). The relative position and orientation of L with respect to P, Y, can be defined in terms of one of the rigid fragments of each molecule. Finally, we use unit quaternions [17] to represent orientation. Since the quaternions q and $-q$ represent the same rotation, orientations are given by an *axis* [18] of the sphere S^3.

The coordinates making up Ξ are then: (a) the position component of Y, a point in $\mathbb{R}^3$; (b) the orientation component of Y, an axis of the sphere S^3; and (c) the dihedral angles of the rotatable bonds of the ligand and protein, points on the circle S^1. The definition of $\langle \Xi \rangle$ and $\delta\Xi$ is straightforward for (a), since we use the normal arithmetic definitions. For (c), we define the mean for each angle [18] by the direction of the mean of the points on S^1 embedded in $\mathbb{R}^2$. We form the deviation in this case by subtraction modulo 2π so that the result lies in $[-\pi, \pi]$.

To find the mean of the orientations (b), we similarly embed S^3 in $\mathbb{R}^4$. We define the mean orientation [18] as the axis in $\mathbb{R}^4$ about which the moment of inertia of the sample axes is minimum. To find the deviation of the orientation, we compute the differential rotation $\delta q = q \langle q \rangle^*$ which takes the mean orientation to the sample orientation and we project this into a "turn" vector $\mathbf{u}$ in the unit ball in $\mathbb{R}^3$ so as to preserve the metric. This is achieved by a generalization of the Lambert azimuthal equal-area projection as described in the Appendix. In this representation, the volume of orientational space (which contributes to the multiplier for the unbound volumes) is $\sigma = \frac{4}{3}\pi$.

Occasionally, the point sampled from w' corresponds to one of the coordinates "wrapping around," i.e., the change in the dihedral lies outside $[-\pi, \pi]$ or the turn $\mathbf{u}$ lies outside the unit ball. In order to preserve detailed balance, we reject the resulting move.

There are four ways in which we can improve the quality of the portals obtained by this method so as to make successful wormhole moves more likely. (1) The Monte Carlo runs used to obtain the samples from which the portals are defined should begin with an "annealing" phase where the temperature is started at some high value and slowly reduced to T at which point we start gathering samples. This allows the Monte Carlo sampling to explore phase space more thoroughly. (2) Other methods of finding conformational energy minima [19] can be applied to provide additional starting points for the Monte Carlo runs. (3) The samples can be supplemented with those obtained by applying those symmetry operations which leave the molecules invariant. In the case of non-chiral ligands, we would also apply a reflection of the ligand in the unbound case since this will leave the energy unchanged. In this way, the portal moves allow all the symmetric variants of the molecules to be explored so that symmetry is included in a systematic way in the computation of the binding affinity. (4) When forming $\langle \delta\Xi_0 \, \delta\Xi_0 \rangle$, we should set the intermolecular cross terms to zero because the conformations of the two molecules are independent when $\lambda = 0$.

This method of finding portals depends on the samples "spanning" a volume of phase space. This requires that the dimensionality of phase space be sufficiently small and this, in turn, implies the use of an implicit solvation model. In addition the portals can be more reliably found if the "hard" degrees of freedom in the molecules are replaced by constraints (e.g., by fixing the bond lengths and bond angles as described above).

The method of successively subdividing the samples may lead to suboptimal portals in some cases, for example, by dividing a contiguous set of samples. We have recently experimented with fitting a mixture of Gaussians to the samples using the expectation-maximization (EM) algorithm [20, 21]. Since this optimizes the fit to all the samples, it usually results in fewer, better, portals. In this case, we are naturally lead to use Gaussians for the portal functions rather than the more restrictive ellipsoids. By enforcing the symmetries of the ligand when making the fits, ligand symmetry can also be included in a rigorous way. These improvements will be described in a subsequent publication.

The free energy calculation

Prior to the calculation of the free energy, we estimate a suitable value of V_0, which enters into the definition of the volumes of the $\lambda = 0$ portals, by assigning an estimated statistical weight of $v \exp(-\beta \langle E^* \rangle)$ to each portal and estimating

$$V_0 \sim \frac{\sum_{\lambda=1} v \exp(-\beta \langle E^* \rangle)}{\sum_{\lambda=0} (v/V_0) \exp(-\beta \langle E^* \rangle)},$$
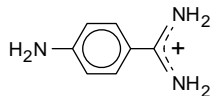
FIG. 2: The structure of p-amino-benzamidine.

where the sums in numerator (denominator) are over the bound (unbound) portals and v/V_0 for the unbound portals is the volume of the ellipsoid (multiplied by σ) and thus does *not* depend on V_0. If this estimate for V_0 results in the sampling being too heavily weighted toward $\lambda = 0$ or 1, V_0 may be changed and the current values of the sample sums for K_d can be adjusted to account for this change.

The choice of the starting configuration for the free energy calculation may introduce some bias in the results. We can remove much of this bias by using the portals to select the starting configuration: select a portal w with probability proportional to its estimated statistical weight; select a configuration, $[\Gamma, \lambda]$, uniformly from this portal; and accept the configuration with probability

$$\min\big[1, \exp\big(-\beta[E_\lambda^*(\Gamma) - \langle E^* \rangle]\big)\big],$$

where $\langle E^* \rangle$ is the canonical average energy over the portal. This procedure is repeated until a configuration is accepted. The bias can be further reduced by running the Monte Carlo calculation for several correlation times, defined by Eq. (6), prior to gathering the data for Eq. (3).

During the course of the free energy calculation, normal and wormhole Monte Carlo moves are mixed. With the normal moves, we only attempt to change Γ and, for this reason, it is possible to have the Monte Carlo step size depend on λ (with, usually, the step size being larger with $\lambda = 0$). In practice, most ($\sim 90\%$) of the attempted moves are wormhole moves because frequently we have $[\Gamma, \lambda] \notin w$ and the attempted wormhole move is inexpensively rejected.

The method is robust in the sense that it does not depend on the particular form of the energy function. In addition, a failure to make transitions between $\lambda = 0$ and 1 can be detected. This may be because the test $[\Gamma, \lambda] \in w$ never succeeds (i.e., the configuration has been trapped in a new well), or because the acceptance criterion is never met, which indicates that there is a deep well within the well of one of the portals. In both cases, the problem can be corrected by adding a new portal based on recent configurations.

Example

The efficacy of the wormhole method depends on how frequently a configuration lies within one of the portals and how often the jump to the new portal is accepted. In order to assess these questions, we have computed the binding affinity of p-amino-benzamidine, whose structure is shown in Fig. 2, to the digestive enzyme trypsin, at $T = 290\,\mathrm{K}$. We emphasize that

the primary goal of this exercise is to assess how well wormhole moves allow the free energy calculation to converge. For this purpose, we are not so interested in comparing the resulting computed binding affinity to the experimental data since this will depend in large measure on the accuracy of the force field and of the protein structure. Nevertheless, since the convergence will depend on the complexity of the energy "landscape," we use this example to epitomize the binding of a small ligand to a protein.

The coordinates of the atoms in trypsin are taken from a trypsin-benzamidine complex, 1BTY [22]. At physiological pH, the amidine group is protonated (net charge of $+1$) and, in the complex, it is attracted to a negatively charged aspartate residue in trypsin inhibiting its enzymatic action. We employ the Amber 7 force field [23, 24, 25] and the GB/SA solvation model [26, 27, 28]. The protein is taken to be rigid, $n_P = 0$, and two bonds of the ligand are allowed to rotate, namely, those connecting the amidine and the amino groups to the benzene ring, $n_L = 2$. The published force field [23] does not provide a satisfactory torsion for the bond between the benzene ring and the amidine group and this term was determined using Gaussian 98 [29] as $[-14.2\cos(2\phi) + 3.3\cos(4\phi) + 0.5\cos(6\phi)]\,\mathrm{kJ/mol}$, where ϕ is the dihedral angle.

Five unbound and five bound canonical simulations of 1000 steps each were carried out to find the portals. The resulting configurations were supplemented by those obtained by applying the symmetry operations which leave the ligand invariant. This gave 16 unbound and 8 bound portals. The configurations of the bound portals are all the same "pose" of the ligand on the protein and correspond to the symmetries of p-amino-benzamidine given by rotating the amino, benzene, and amidine groups by $180°$. The unbound ligand also exhibits 8 symmetries given by rotating the amino and amidine groups by $180°$ relative to the central benzene ring and by including the mirror images. However since the bond parameters allow partially free rotation of the amidine group, each symmetric set of configurations is represented by two portals.

We estimated a suitable value for V_0 of $0.39 \times 10^{-18}\,\mathrm{m}^3$ based on the volumes and the mean energies of the computed portals. During the binding affinity calculation, wormhole moves were attempted on 90% of the steps (the other 10% were standard Monte Carlo moves). Of these attempted wormhole moves, 97% failed (inexpensively) because the configuration was not in the chosen w. Of the remaining 3%, about 60% lead to a successful move, of which about 40% involved switching from $\lambda = 0$ to 1 and *vice versa*. The major computational cost in the free energy calculation is the evaluation of the bound energies. In this example, the wormhole moves required less than 3 bound energy evaluations, on average, to effect the transition from a bound configuration to an unbound one and back. This enables an accurate estimate to be made of the ratio of the averages in Eq. (3) which after 5×10^6 steps yields $\mathrm{p}K_d = 7.99 \pm 0.01$. The error estimate is derived in the next section and represents a 2% relative error in K_d.
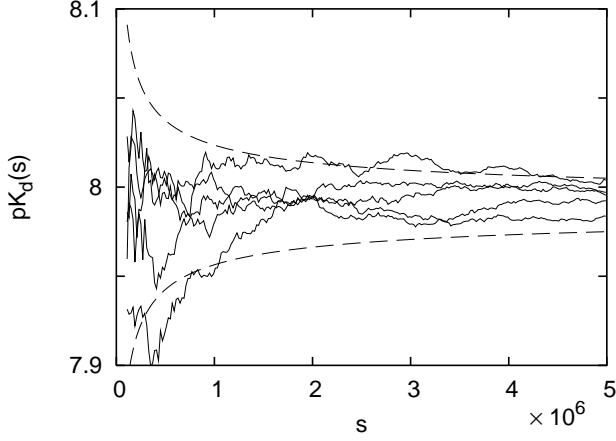
FIG. 3: Cumulative estimates $pK_d(s)$ obtained by sampling every 100th step from the first s steps of 5 independent Monte Carlo runs. The dashed lines shows convergence as $1/\sqrt{s}$ to the mean value of 7.99.

This example illustrates that the method is effective at allowing sufficient transitions between the bound and unbound states to enable the binding affinity of protein-ligand systems to be accurately computed. We note that our computed binding affinity differs from the experimental results of 5.1 to 5.2 [30, 31]. This discrepancy may be accounted for by modest, $\sim 20\%$, errors in the force field.

Error analysis

In order to assess the errors in the computation of pK_d in more detail, we carried out 10 independent runs similar to the one described in previous section. Each of these used the same portals and the same value of V_0. We computed cumulative estimates for pK_d based on the first s steps of the Markov chains. When forming the averages in Eq. (3) we sample every 100th step. (As we shall see, there is a high degree of correlation within 100 steps; thus there would be little improvement in the estimate of pK_d by sampling more frequently.) The results for 5 of these runs are shown in Fig. 3. The convergence toward the mean is as $1/\sqrt{s}$; after 5×10^6 steps, the error in pK_d has been reduced to about ± 0.01.

For the purposes of further analysis, let us assume that the computation is carried out with $E_\lambda^*(\Gamma) = E_\lambda(\Gamma)$ so that all samples have the same statistical weight. The probability that the system is in the bound (resp. unbound) state is $p = \langle \delta_{\lambda 1} \rangle$ (resp. $q = 1 - p = \langle \delta_{\lambda 0} \rangle$) and Eq. (3) becomes

$$K_d = \frac{1}{V_0} \frac{q}{p}. \qquad (5)$$

The determination of K_d is then equivalent to estimating q/p by taking the ratio of the number of unbound and bound steps in the Monte Carlo simulation. How well this estimate con-
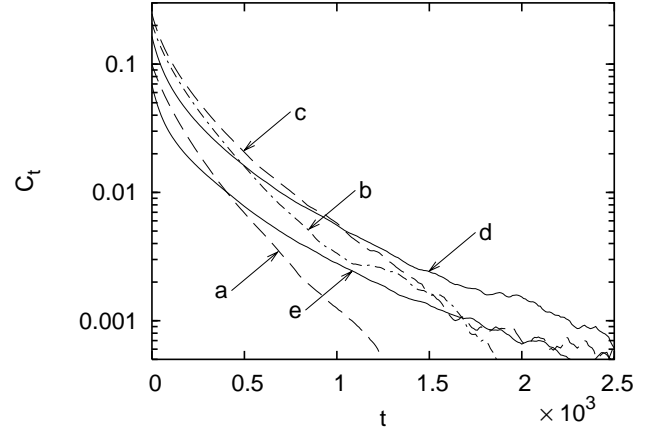

FIG. 4: The λ-correlation function C_t. The curves show C_t for $V_0/(0.39 \times 10^{-18} \text{ m}^3) =$ (a) $\frac{1}{10}$, (b) $\frac{1}{3}$, (c) 1, (d) 3, and (e) 10. The ratio q/p varies correspondingly (from approximately $\frac{1}{8}$ to $\frac{100}{8}$). The correlation times are (a) 329, (b) 346, (c) 352, (d) 382, and (e) 415. In determining the correlation times, we limit the sum in Eq. (6) to $0 < t \leq 2500$ in order to avoid including the small but noisy terms for large t.

verges depends, naturally, on the "correlation time" of the system. If wormhole moves rarely cause the system to switch between bound and unbound states, the correlation time is large and the convergence will be slow.

In order to make these ideas quantitative, we define the λ-correlation function,

$$C_t = \langle (\lambda_s - p)(\lambda_{s+t} - p) \rangle_s,$$

where λ_s is the value of λ at simulation step s and $\langle \ldots \rangle_s$ denotes an average over steps. Figure 4 shows C_t for several different values of V_0. (From Eq. (5), we have $q/p \propto V_0$.) In a Markov chain of length s, the expected number of bound states is ps, while, for $s \to \infty$, the variance in the number of bound states is $2Ds$, where

$$D = \frac{1}{2} C_0 + \sum_{t>0} C_t = \frac{1}{2} C_0 \tau. \qquad (6)$$

This provides us with the definition of the correlation time, τ. From Fig. 4, we see that C_t decays approximately exponentially so that the sum converges.

We can compare the Monte Carlo simulation to a simpler Markov process of independent trials, e.g., tossing a coin where the probability of heads is p. In this case, we have $C_{t>0} = 0$ and $D = \frac{1}{2} C_0 = \frac{1}{2} pq$. In the limit $s \to \infty$, the relative errors in estimating p for the Monte Carlo simulation will be the same as that for s/τ tosses of the coin. We can illustrate this by making 5 independent simulations of a coin tossing experiment to match the data in Fig. 3, for which $p = 0.443$ and $\tau = 352$. The results are given in Fig. 5, where we plot h_n/t_n against n where h_n (resp. t_n) is the number of
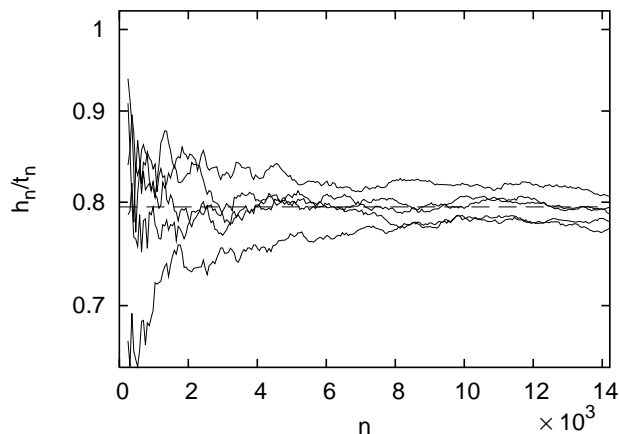
FIG. 5: Estimation of the bias of a coin with $p = 0.443$. Five independent runs are shown. In each run, the coin was tossed $5 \times 10^6/\tau = 14200$ times, where τ is the correlation time for Fig. 4(c), and the cumulative value of h_n/t_n is recorded. The expected value, p/q, is shown as a dashed line. The axes have been adjusted to allow the figure to be directly compared with Fig. 3.

heads (resp. tails) after n throws. As expected, Figs. 3 and 5 both exhibit the same convergence behavior.

Since the distribution of outcomes h_n in the coin tossing experiment is the binomial distribution, the mean and variance for $l_n = \ln(h_n/t_n)$ can be calculated, yielding

$$\langle l_n \rangle = \ln\left(\frac{p}{q}\right) + \frac{1}{2n}\left(\frac{p}{q} - \frac{q}{p}\right) + O(n^{-2}),$$

$$\langle (l_n - \langle l_n \rangle)^2 \rangle = \frac{1}{npq} + O(n^{-2}).$$

The standard error in the estimate of pK_d from a Monte Carlo run with s steps is found by substituting $n = s/\tau$ and scaling by $\ln 10$ (since pK_d is defined in terms of the common logarithm) to give

$$\Delta pK_d \approx \frac{1}{\ln 10}\sqrt{\frac{\tau}{spq}}. \tag{7}$$

In the case of the simulations shown in Fig. 3, we find $\Delta pK_d \approx 0.01$ consistent with the data in the figure. From Fig. 4, we see that τ is weakly dependent on p. Thus for a given s, we minimize the error by choosing $p = q = \frac{1}{2}$. Alternatively, we may wish to minimize the error for a given amount of computational effort. If bound steps are f times more expensive to carry out than unbound steps, the error is minimized by taking $q/p = \sqrt{f}$, i.e., $p = 1/(1 + \sqrt{f})$.

Discussion

The important aspects of this method that enable us to compute the binding affinity of a ligand to a protein are: (1) formulation in the extended $[\Gamma, \lambda]$ space, which allows the unbound and bound systems to be treated as a single canonical system and K_d to be expressed as the ratio of canonical averages, Eq. (3); (2) the wormhole move which allows transitions between the bound and unbound systems; and (3) the use of an implicit solvation model which reduces the number of degrees of freedom and so allows portals to be identified.

A method similar to ours is "simulated mutational equilibration," [32] which employs an extended phase space, uses an implicit solvation model, and allows jumps between different systems. However, in place of the wormhole move, this work employed a more restrictive "jumping between wells" move, which limited its applicability to computing the difference in the binding affinities for two enantiomers. In contrast, our wormhole method can be used to compute directly the binding affinity of a ligand with several rotatable bonds to a protein target. This allows us to study a wide range of interesting drug-like ligands.

In the special case of binding rigid molecules (for which we have $E_0^*(\Gamma) = 0$), wormhole Monte Carlo is isomorphic to a grand canonical Monte Carlo simulation [33]. Consider a grand canonical system with a single protein molecule and with ligand molecules at a fixed chemical potential μ; we make the additional restriction that the ligand-ligand interaction energy is infinite, so that the system can only accommodate 0 or 1 ligand molecule. For the wormhole simulation, we specify the bound portal to include the full simulation volume with arbitrary orientation; the unbound port is degenerate and corresponds merely to specifying $\lambda = 0$. The wormhole moves from $\lambda = 0$ to 1 and *vice versa* correspond to particle insertion and deletion in the grand canonical simulation; and we can verify the acceptance probabilities are the same.

The application described here can be extended by allowing a greater degree of flexibility for the protein and the ligand. This permits the treatment of side-chain rotation and ligand-induced loop movement on the part of the protein and the treatment of flexible rings for the ligand. Many standard Monte Carlo techniques can be used with this method, if appropriate—preferential sampling, early rejection, force bias, etc. Wormhole moves could also be used to treat other discrete, or nearly discrete, transitions. Examples are: the treatment of molecules, such as cyclohexane, which can assume distinct conformations; discrete sets of side chain rotations in the protein; protonation and tautomerization states for either molecule. In each of these cases, the acceptance probability for the transitions should account for the free energy difference between the discrete states.

Appendix: Generalized Lambert projection

The Lambert azimuthal equal-area projection projects a point on S^2 to a point on a disk in $\mathbb{R}^2$. Here, we will generalize this to arbitrary dimensions, i.e., we will find a projection from S^n to a ball in $\mathbb{R}^n$. (Thus the circle S^1 is projected into a line segment; the surface of a sphere S^2 is project into a disk; S^3 is projected into a sphere; etc.) The projection is azimuthal, so that directions from the pole are preserved in the projected space.

The sequence S^n can be defined recursively by

$$
\begin{aligned}
S^0 &= [\pm 1], \\
S^n &= [\cos\theta, \sin\theta\, S^{n-1}],
\end{aligned}
$$

where θ is the colatitude and $0 \le \theta \le \pi$. The area of S^n lying between θ and $\theta + d\theta$ is therefore

$$
dA = a_{n-1} \sin^{n-1}\theta\, d\theta,
$$

where a_n is the area of S^n. We project that portion of S^n with colatitude in $[0, \theta]$ to a ball in $\mathbb{R}^n$ of radius t. Equating the measures (area on S^n and volume in $\mathbb{R}^n$), we obtain

$$
v_n t^n = a_{n-1} \int_0^\theta \sin^{n-1}\theta'\, d\theta'
$$

where v_n is the volume of a unit ball in $\mathbb{R}^n$. Using the relation,

$$
a_{n-1} = \left.\frac{d(v_n t^n)}{dt}\right|_{t=1} = n v_n,
$$

we obtain

$$
t = \left(n \int_0^\theta \sin^{n-1}\theta'\, d\theta' \right)^{1/n}. \tag{8}
$$

The general "equal-area" projection is given by the mapping

$$
[\cos\theta, \sin\theta\, S^{n-1}] \to \mathbf{t},
$$

where $\mathbf{t} \in \mathbb{R}^n$, $\hat{\mathbf{t}}$ is given by the point on S^{n-1}, and t is given by Eq. (8). Some special cases of Eq. (8) are

$$
t = \begin{cases}
\theta, & \text{for } n = 1, \\
2\sin\frac{1}{2}\theta, & \text{for } n = 2, \\
\left[\frac{3}{4}(2\theta - \sin 2\theta)\right]^{1/3}, & \text{for } n = 3, \\
2\sin\frac{1}{2}\theta\left[\frac{1}{3}(1 + 2\cos^2\frac{1}{2}\theta)\right]^{1/4}, & \text{for } n = 4.
\end{cases}
$$

The case $n = 1$ corresponds to unwrapping a circle onto a line; and $n = 2$ gives the Lambert azimuthal equal-area projection. We are interested in the case $n = 3$ as a way of mapping orientations from unit quaternion space to $\mathbb{R}^3$. A quaternion $q = [\cos\theta, \sin\theta\,\hat{\mathbf{u}}]$ represents a rotation of $\psi = 2\theta$ about $\hat{\mathbf{u}}$. Noting that q and $-q$ represent the same rotation, we may make the restriction $0 \le \theta \le \frac{1}{2}\pi$. We choose, therefore, to map this hemisphere of S^3 onto $\mathbf{u}$ in the unit ball, by making the substitutions $\theta = \frac{1}{2}\psi$ and $t = \left(\frac{3}{4}\pi\right)^{1/3}u$ to give

$$
u = \left(\frac{\psi - \sin\psi}{\pi} \right)^{1/3}.
$$

(The other hemisphere, corresponding to $\frac{1}{2}\pi < \theta \le \pi$, maps onto the shell $1 < u \le 2^{1/3}$.) We call this $\mathbf{u}$ representation of orientations "turn space."

---

[*] Electronic address: ckarney@sarnoff.com

[†] Current address: CRPP, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland.

[1] M. Mezei and D. L. Beveridge. Free energy simulations. *Ann. N.Y. Acad. Sci.*, **1986**, *482*, 1–23.

[2] P. A. Kollman. Free energy calculations: Applications to chemical and biochemical phenomena. *Chem. Rev.*, **1993**, *93*, 2395–2417.

[3] L. D. Landau and E. M. Lifshitz. *Statistical Physics*, volume 5 of *Course of Theoretical Physics*. Pergamon Press, 2nd edition, 1969.

[4] C. H. Bennett. Efficient estimation of free energy differences from Monte Carlo data. *J. Comp. Phys.*, **1976**, *22*, 245–268.

[5] H. Luo and K. Sharp. On the calculation of absolute macromolecular binding free energies. *Proc. Nat. Acad. Sci.*, **2002**, *99*, 10399–10404.

[6] M. K. Gilson, J. A. Given, B. L. Bush, and J. A. McCammon. The statistical-thermodynamic basis for computation of binding affinities: A critical review. *Biophys. J.*, **1997**, *72*, 1047–1069.

[7] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, **1953**, *21*, 1087–1092.

[8] A. P. Lyubartsev, A. A. Martsinovski, S. V. Shevkunov, and P. N. Vorontsov-Velyaminov. New approach to Monte Carlo calculation of the free energy: Method of expanded ensembles. *J. Chem. Phys.*, **1992**, *96*, 1776–1783.

[9] B. Tidor. Simulated annealing on free energy surfaces by a combined molecular dynamics and Monte Carlo approach. *J. Phys. Chem.*, **1993**, *97*, 1069–1073.

[10] X. Kong and C. L. Brooks III. λ-dynamics: A new approach to free energy calculations. *J. Chem. Phys.*, **1996**, *105*, 2414–2423.

[11] The term "wormhole" is employed to evoke the idea of interstellar travel via wormholes [M. S. Morris, K. S. Thorne, and U. Yurtsever, *Phys. Rev. Lett.*, **1988**, *61*, 1446–1449]. This wormhole move is not to be confused with a different move of the same name used in polymer simulations [J. Houdayer, *J. Chem. Phys.*, **2002**, *116*, 1783–1787].

[12] M. A. Miller and W. P. Reinhardt. Efficient free energy calculations by variationally optimized metric scaling: Concepts and applications to the volume dependence of cluster free energies and to solid-solid phase transitions. *J. Chem. Phys.*, **2000**, *113*, 7035–7046.

[13] Z. Zhu, M. E. Tuckerman, S. O. Samuelson, and G. J. Martyna. Using novel variable transformations to enhance conformational sampling in molecular dynamics. *Phys. Rev. Lett.*, **2002**, *88*, 100201.

[14] A. F. Voter. A Monte Carlo method for determining free-energy differences and transition state theory rate constants. *J. Chem. Phys.*, **1985**, *82*, 1890–1899.

[15] H. Senderowitz, F. Guarnieri, and W. C. Still. A smart Monte Carlo technique for free energy simulations of multiconformational molecules: Direct calculations of the conformational populations of organic molecules. *J. Am. Chem. Soc.*, **1995**, *117*, 8211–8219.

[16] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns

Hopkins University Press, 3rd edition, 1996.

[17] W. R. Hamilton. On quaternions. *Proc. Royal Irish Acad.*, **1847**, *3*, 1–16. URL http://www.maths.tcd.ie/pub/HistMath/People/Hamilton/Quatern2/.

[18] K. V. Mardia and P. E. Jupp. *Directional Statistics*. J. Wiley & Sons, 1999.

[19] M. S. Head, J. A. Given, and M. K. Gilson. Mining minima: Direct computation of conformational free energy. *J. Phys. Chem.*, **1997**, *101*, 1609–1618.

[20] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Stat. Soc.*, **1977**, *39*, 1–38.

[21] J. J. Verbeek, N. Vlassis, and B. Kröse. Efficient greedy learning of Gaussian mixture models. *Neural Computation*, **2003**, *15*, 469–485.

[22] B. A. Katz, J. Finer-Moore, R. Mortezaei, D. H. Rich, and R. M. Stroud. Episelection: Novel $K_i \sim$ nanomolar inhibitors of serine proteases selected by binding or chemistry on an enzyme surface. *Biochemistry*, **1995**, *34*, 8264–8280. URL http://www.rcsb.org/pdb/cgi/explore.cgi?pdbId=1bty.

[23] D. A. Case, D. A. Pearlman, J. W. Caldwell, T. E. Cheatham, III, J. Wang, W. S. Ross, C. L. Simmerling, T. A. Darden, K. M. Merz, R. V. Stanton, *et al. Amber 7*. University of California, San Francisco, 2002. URL http://amber.scripps.edu/doc7/.

[24] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, Jr., D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. A second generation force field for the simulation of proteins, nucleic acids and organic molecules. *J. Am. Chem. Soc.*, **1995**, *117*, 5179–5197.

[25] C. I. Bayley, P. Cieplak, W. D. Cornell, and P. A. Kollman. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: The RESP method. *J. Phys. Chem.*, **1993**, *97*, 10269–10280.

[26] W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.*, **1990**, *112*, 6127–6129.

[27] G. D. Hawkins, C. J. Cramer, and D. G. Truhlar. Pairwise solute descreening of solute charges from a dielectric medium. *Chem. Phys. Lett*, **1995**, *246*, 122–129.

[28] V. Tsui and D. A. Case. Molecular dynamics simulations of nucleic acids with a generalized Born solvation model. *J. Am. Chem. Soc.*, **2000**, *122*, 2489–2498.

[29] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, V. G. Zakrzewski, J. A. Montgomery, Jr., R. E. Stratmann, J. C. Burant, *et al. Gaussian 98, Revision A.9*. Gaussian, Inc., Pittsburg, 1998. URL http://www.gaussian.com.

[30] M. Mares-Guia and E. Shaw. Studies of the active center of trypsin. *J. Biol. Chem.*, **1965**, *240*, 1579–1585.

[31] S. M. Schwarzl, T. B. Tschopp, J. C. Smith, and S. Fischer. Can the calculation of ligand binding free energies be improved with continuum solvent electrostatics and an ideal-gas entropy correction? *J. Comp. Chem.*, **2002**, *23*, 1143–1149.

[32] H. Senderowitz, D. Q. McDonald, and W. C. Still. A practical method for calculating relative free energies of binding: Chiral recognition of peptidic ammonium ions by synthetic ionophores. *J. Org. Chem.*, **1997**, *62*, 9123–9127.

[33] D. J. Adams. Grand canonical ensemble Monte Carlo for a Lennard-Jones fluids. *Mol. Phys.*, **1975**, *29*, 307–311.